Comparative genomics of *Escherichia coli* causing bloodstream infections in the Hospital for Tropical Diseases, Ho Chi Minh city, Vietnam

Tu Thi Phuong Le

Wellcome Trust Sanger Institute

University of Cambridge

Darwin College

This dissertation is submitted for the degree of Master of Philosophy in Biological Science

September 2017

Declaration

I hereby declare that:

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text and bibliography. My dissertation is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other University. I further state that no part of my dissertation has already been or is being concurrently submitted for any such degree, diploma or other qualification.

I confirmed that my dissertation does not exceed the limit of length of 20,000 words prescribed in the Special Regulations of the MPhil examination for which I am a candidate.

Acknowledgements

This thesis was possible because of the collective efforts of many people, from the patient beds to lab benches and computer clusters, trying to deconvolute one of the oldest and cheekiest bacteria that cause big issues not only in a single referral hospital in Vietnam, but also across the globe.

This has been a tough but wonderful journey. I was extremely grateful to work on this project and have seen myself grow academically, intellectually, and personally in the past year in Cambridge.

First of all, I would like to thank my main supervisor Prof. Nicholas Robert Thomson for his unbelievable patience and guidance during the research. His continuous encouragement and support during difficult times is something that I always greatly appreciated.

I also would like to thank my former supervisor Prof. Stephen Baker from Oxford University Clinical Research Unit (OUCRU) for coming up with the idea for this project during the application process to the MPhil program as it turned out to be a very interesting project.

I would like to express my gratitude to my thesis committee members including Julian Rayner, Stephen Bentley, Estee Torok, and Kate Baker, who have provided valuable feedback and criticism.

I would like to thank members of Team 216 at the Sanger Institute for their tremendous daily support, as well as for providing me with a remarkably happy and friendly place to work. In no particular order, these include Angèle Bénard, Mathew Beale and Déborah Kreyenbihler. With no prior background on working with whole genome sequencing data, I therefore could not have completed the project without the guidance and patience of Alison Mather (Cambridge University), Daryl Domman, Eva Heinz and Alex Wailan (Sanger Institute), who never hesitating to spend time teaching me when I was stuck. I also would like to thank Ngure Kagia, another MSc student, who helped me to feel comfortable using Linux and computer clusters at the very beginning of my time at Sanger. Finally, I am indebted to Matthew Dorman for giving very constructive helpful feedback during my writing process.

The blood-derived *Escherichia coli* strains from this six-year period is the effort of all Microbiology staff at Hospital for Tropical Disease (Ho Chi Minh city, Vietnam), especially Nguyen Huu Hien and Tran Thi Ngan who carried out the blood culturing, identification, antimicrobial phenotype testing and sample archiving. The clinical metadata extracted from electronic hospital record database for this study was kindly provided by Nguyen Huu Hien, Dr Nguyen Phu Huong Lan and Dr Nguyen Van Vinh Chau. The second set of blood and stool isolates, as well as its associated clinical data, came from a big study on hospital acquired infections in the Intensive Care Unit lead by Dr Duong Bich Thuy, Mr James Campell, Dr Louise Thwaites, Prof Stephen Baker and Prof Guy Thwaites. I also would like to thank Nguyen Hoang Thu Trang, To Thi Thanh Huyen and Vong Vinh Phat for assisting me during the DNA extraction stage. I also would like to thank Dr Ho Dang Trung Nghia for discussions about the context of management and treatment of BSI patients in the hospital. My work relied on enormous sequencing

data of nearly 700 bacterial isolates, therefore I am indebted to the DNA core sequencing team and also the Pathogen Informatics team at the Sanger Institute for sample management, sequencing, and providing many analysis pipelines and bioinformatics supports. Many scripts and software I used for WGS analysis in this thesis were written by Simon Harris and Andrew Page.

Apart from working, I would like to give special thanks to Benjamin Bai, Simon Scutts, other friends and colleagues at Sanger and Darwin college badminton club for keeping me sane through many fun and joyful badminton matches.

Finally, this study could not have made possible without the support of the Wellcome Trust, who generously funded not only my university fees and living expenses, but also the research costs in both institutions in two nations, Sanger (UK) and OUCRU (Vietnam). I hope my study will shed a light on how the cheeky *E. coli* cause disease in Vietnamese patients as well as acquire drug resistance over time. Future work will be required to understand how to prevent and reduce the burden of disease in Vietnamese patients.

To my parents and little sister,

for their endless and unconditional love.

Table of Contents

Figures

Tables

Abstract

Escherichia coli (*E. coli*) is a versatile bacterium, with the capability to act not only as a commensal coloniser but also as a pathogen that can cause invasive disease. Currently, *E. coli* is the leading cause of bloodstream infections in both developed and developing countries, accounting for 25-30% of bacteraemia cases globally. *E. coli* bacteraemia is further exacerbated by the emergence of antimicrobial resistance, particularly extended spectrum β-lactamases (ESBLs) and carbapenemases in Gram-negative bacteria. Understanding the nature and diversity of *E. coli* causing bloodstream infections is crucial for the enhancement of infection control measures, to minimise the further emergence of antimicrobial resistant isolates, and for the reduction of morbidity and mortality.

We used whole-genome sequencing to analyse the population structure of 506 invasive and 159 carriage *E. coli* isolates collected at The Hospital for Tropical Diseases, Vietnam during 2010-2015, and to look for genetic signatures that differentiate them from one another. We found substantial diversity therein, both in the total number of sequence types present, and in the number of resistant genes carried. Among blood isolates, ST131, ST95, ST69, ST1193 and ST73 were dominant STs, while from the rectal swab (carriage) isolates, ST131, ST1193 and ST648 were most prevalent. All of these STs from blood and rectal swab samples remained dominant over the entire study period, the exception being ST1193. Interestingly ST1193 was not present in Vietnam prior to 2011, but once introduced, it quickly emerged and replaced other more drug sensitive *E. coli* clones. From longitudinal sampling and paired *E. coli* isolated from blood and rectal swab from the same patients, we demonstrated that majority of *E. coli* infections (57%) were acquired from the patients' own gut microbiota. This study shows that several genetic factors, including genes that mediate adhesion (fimbriae and pili), iron acquisition (siderophores), immune evasion (capsule synthesis) and toxin elaboration (haemolysin and cytotoxic necrotizing factor 1) are significantly more common in invasive *E. coli* than in carriage strains. We were also able to begin to explain anomalies in the patterns of antimicrobial resistance of blood and rectal swab *E. coli* isolates using the population structure we defined here.

Taken together, we have generated a genetic framework for future studies focusing on *E. coli* BSI at HTD. These data also demonstrate that a combination of both virulence genes and antimicrobial resistance genes are essential for the success of certain lineages in causing invasive disease in Vietnamese patients. Perhaps more importantly we also show that patients who develop bacteraemia in hospital did so largely through infection with isolates already present in their intestinal tract and so were predominantly community acquired invasive infections rather than hospital acquired infections, contrary to our original hypothesis.